LOCAL PHASE U-NET FOR FUNDUS IMAGE SEGMENTATION

Sudhakar Kumawat and Shanmuganathan Raman

Computer Science and Engineering, Indian Institute of Technology Gandhinagar, India
{sudhakar.kumawat, shanmuga}@iitgn.ac.in

ABSTRACT

In this paper, we propose Rectified Local Phase Unit (ReLPU), which is an efficient and trainable convolutional layer that utilizes phase information computed locally in a window for every pixel location of the input image. The ReLPU layer is based on applying the Rectified Linear Unit (ReLU) activation function on the local phase information extracted by computing the local Fourier transform of the input image at multiple low frequency points. The ReLPU layer, when used at the top of the segmentation network U-Net, is observed to improve the performance of the baseline U-Net model. We demonstrate this using the task of segmenting blood vessels in fundus images of two standard datasets, DRIVE and STARE, achieving state-of-the-art results. An important feature of the ReLPU layer is that it is trainable which allows it to choose the best frequency points for computing local Fourier transform and to selectively give more weight to them during training.

Index Terms— Convolutional neural network, phase information, fundus image segmentation.

1. INTRODUCTION

Over the past few years, with the availability of large-scale datasets and computation power, deep learning has achieved impressive results in a wide range of applications in the fields of computer vision, artificial intelligence, and image processing. In fact, in a majority of the computer vision problems like image classification, semantic segmentation, object detection and many more, development in neural network architectures like Convolutional Neural Networks (CNNs) have achieved state-of-the-art results. Each year, new training methods and network architectures are being introduced with the goal of developing models that can represent the underlying data for a given task in the best possible way.

In this paper, we propose a phase-based learnable convolutional layer named Rectified Local Phase Unit (ReLPU). The ReLPU layer when used at the top (just after the input layer) of the segmentation network U-Net, such that the input to the network is the local phase information of the input images, is observed to significantly improve the performance of the baseline U-Net model. We demonstrate this using the task of segmenting blood vessels in the fundus images of two standard datasets, DRIVE and STARE, achieving state-of-the-art results on both. Our results show that the local phase information of the input image provides a better representation capability than the spatial information that is generally used by most CNNs. The ReLPU layer is based on applying the Rectified Linear Unit (ReLU) activation function on the phase information of the local Fourier transform at multiple low frequency points of the input. The ReLPU layer is trainable, which allows it to choose the optimal frequency points and to selectively give more weight to them during training.

S. Kumawat was supported by TCS Research Fellowship and S. Raman was supported by a SERB Core Research Grant.

2. RELATED WORK

The Short Term Fourier Transform (STFT) in 2D space was first studied by Hinman *et al.* in [1] as an efficient tool for image encoding. The 2D STFT has three features which made it an attractive technique for image coding. They are: its excellent energy compaction, its ability to decorrelate the input features, and that it is free of the "blocking effects" [1]. Natural images are often composed of objects with sharp edge features. It has been observed that Fourier phase information is more important in accurately representing these edge features than the magnitude information. Since 2D STFT is simply a windowed Fourier transform, the same property applies. Ojansivu and Heikkilä proposed the Local Phase Quantization (LPQ) operator for blur invariant texture analysis [2]. The LPQ operator is based on the binary encoding of the phase information of the local Fourier transform at low frequency points. Local phase from 1D STFT has been explored recently in deep neural networks such as Fully Complex-valued Deep Neural Network (FCDNN) in [3] for speech processing. To the best of our knowledge, the local phase information extracted from 2D STFT has not been studied in the domain of CNNs for image processing and computer vision applications.

3. METHOD

The ReLPU layer utilizes the phase information computed locally in a window at every position of the input image. It

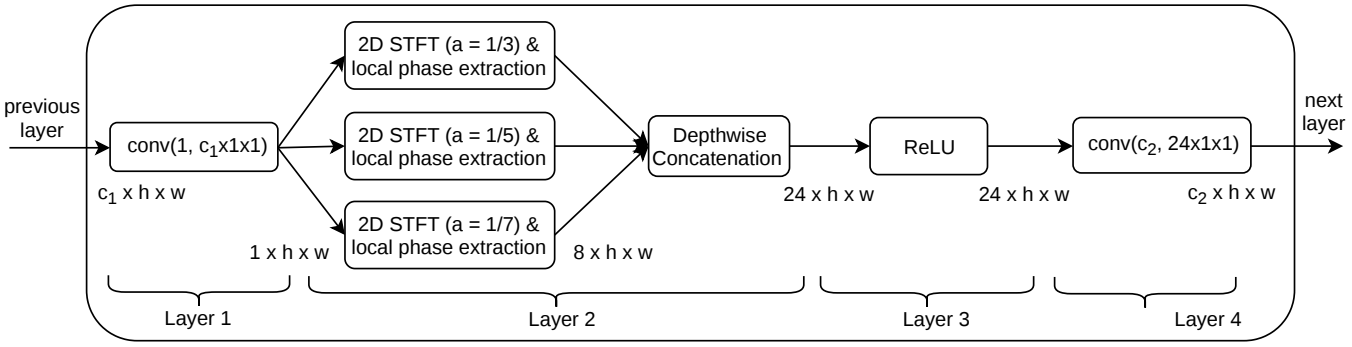


Fig. 1: The architecture of the proposed ReLPU layer.

is a four-layer alternative representation of the standard 2D convolutional layer. Fig. 1 illustrates the architecture of the ReLPU layer. The *first* layer is the standard 1×1 trainable convolution layer containing a single filter of depth c_1 which takes an input image of size $c_1 \times h \times w$ from the previous layer and converts it into a single channel output of size $1 \times h \times w$, where c_1 is the number of channels and $h \times w$ is the spatial dimension of the image. If the image is a gray-scale image, then this layer is optional. Let $f(\mathbf{x})$ be the output of the *first* layer. In further discussion, we will use the term "tensor" for the inputs and the outputs of the intermediate layers.

The *second* layer extracts the local phase spectra of $f(\mathbf{x})$ by computing the 2D STFT in the local $M \times M$ neighborhood $\mathcal{N}_{\mathbf{x}}$ of each position $\mathbf{x}$ of the input $f(\mathbf{x})$ using Equation (1).

$$F(\mathbf{u}, \mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{N}_{\mathbf{x}}} f(\mathbf{x} - \mathbf{y}) \exp^{-j2\pi \mathbf{u}^T \mathbf{y}} \tag{1}$$

Here, $\mathbf{u}$ is a 2D frequency variable and $j = \sqrt{-1}$. Using vector notation, we can rewrite Equation (1) as shown in Equation (2).

$$F(\mathbf{u}, \mathbf{x}) = \mathbf{w}_{\mathbf{u}}^T \mathbf{f}_{\mathbf{x}} \tag{2}$$

Here, $\mathbf{w}_{\mathbf{u}}$ is the basis vector of the 2D STFT at frequency $\mathbf{u}$ and $\mathbf{f}_{\mathbf{x}}$ is a vector containing all positions from the neighborhood $\mathcal{N}_{\mathbf{x}}$. Note that, due to the separability of the basis functions, 2D STFT can be efficiently computed for all the positions $\mathbf{x}$ in $f(\mathbf{x})$ by using simple 1D convolutions for the rows and the columns, successively.

In this work, four 2D frequency variables are considered - $\mathbf{u}_1 = [a, 0]^T, \mathbf{u}_2 = [0, a]^T, \mathbf{u}_3 = [a, a]^T$, and $\mathbf{u}_4 = [a, -a]^T$, where a is a scalar such that $a = 1/M$. Let

$$\mathbf{W} = [\Re\{\mathbf{w}_{\mathbf{u}_1}, \mathbf{w}_{\mathbf{u}_2}, \mathbf{w}_{\mathbf{u}_3}, \mathbf{w}_{\mathbf{u}_4}\}, \Im\{\mathbf{w}_{\mathbf{u}_1}, \mathbf{w}_{\mathbf{u}_2}, \mathbf{w}_{\mathbf{u}_3}, \mathbf{w}_{\mathbf{u}_4}\}]^T \tag{3}$$

Here, $\mathbf{W}$ is the $8 \times M^2$ transformation matrix containing all the basis vectors corresponding to the four 2D frequency variables. $\Re\{\cdot\}$ and $\Im\{\cdot\}$ return the real and the imaginary parts of a complex number, respectively. Hence, from Equation (2) and (3) the vector form of the 2D STFT for all the four frequencies $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$ and $\mathbf{u}_4$ can be written as shown in Equation (4).

$$\mathbf{F}_{\mathbf{x}} = \mathbf{W}\mathbf{f}_{\mathbf{x}} \tag{4}$$

Note that, $\mathbf{F}_{\mathbf{x}}$ is computed for all positions $\mathbf{x}$ of the input $f(\mathbf{x})$, resulting in an output with size $8 \times h \times w$.

Note that in the *second* layer in Fig. 1, the phase is extracted by computing 2D STFT in parallel over multiple neighborhood sizes ($M = 3, 5, 7$) and frequency points ($a = 1/3, 1/5, 1/7$) and then concatenated channel-wise. The basic idea behind this approach is to let the network learn on its own what neighborhood sizes and frequency points to choose from and give more weight to the selected ones in the *fourth* layer which is a trainable layer.

The *third* layer is the ReLU activation layer which takes as input a tensor of size $24 \times h \times w$ from the *second* layer and outputs a tensor of the same size.

The *fourth* layer is the standard trainable 1×1 convolutional layer containing c_2 filters, each one of them has a depth equal to 24 which takes as input a tensor of size $24 \times h \times w$ and outputs a tensor of size $c_2 \times h \times w$. This layer can be interpreted as c_2 linear combinations of the channels of the input. Note that it is this layer that gets learned during the training phase of the CNN.

Parameter analysis of the ReLPU layer. The ReLPU layer uses significantly less learnable parameters when compared to the standard convolutional layer with a particular size of the filters and a particular number of input-output channels. Consider a standard convolutional layer with c_1 input and c_2 output channels. Let $p \times p$ be the size of the filters. Thus, the total number of learnable parameters in a standard convolutional layer is $c_1 \cdot p \cdot p \cdot c_2$. An ReLPU layer with c_1 input channels and c_2 output channels consists of just $c_2 \cdot 24 + c_1 \cdot 1$ learnable parameters. Thus, the ratio of the number of learnable parameters in a standard convolutional layer and the ReLPU layer is given by:

$$\frac{\text{No. of params. in CNN layer}}{\text{No. of params. in ReLPU layer}} = \frac{c_1 \cdot p \cdot p \cdot c_2}{c_2 \cdot 24 + c_1 \cdot 1}$$

For simplicity, assume $c_2 = c_1$ and $c_1 = 25$. This reduces the above ratio to p^2. Thus, for a filter of size 3×3 in the standard convolutional layer, the ReLPU layer uses nine times less learnable filters. Therefore, numerically, ReLPU layer saves atleast $9\times$, $25\times$, $49\times$, $81\times$, $121\times$, and $169\times$ parameters during learning for 3×3, 5×5, 7×7, 9×9, 11×11, and 13×13 convolutional filters, respectively.

Statistical advantages of the ReLPU layer. An important advantage of using the ReLPU layer is that it decorrelates the input features due to its use of the 2D STFT which is known to have such a property [1]. Recent works on regularizing CNNs such as [4, 5] have shown that the decorrelation of features enables us to achieve better performance.

4. EXPERIMENTS

4.1. Datasets

We test the proposed ReLPU layer on the task of segmenting blood vessels in fundus images. For this, we use two standard publicly available datasets of fundus images: DRIVE [6] and STARE [7]. The DRIVE dataset consists of 40 eye-fundus color images taken with a Canon CR5 non-mydriatic 3CCD camera with a 45° field of view (FOV), 8 bits per color channel, and at a resolution of 565×584 pixels. These images are further partitioned into a training and a testing set with 20 images in each set. Each of the images in the training set have a manual annotation associated with it while two manual annotations per image are available for the test images. The STARE dataset consists of 20 fundus color images captured with a TopCon TRV-50 fundus camera with 35° FOV, 8 bits per color channel, and at a resolution of 700×605 pixels. Each image is manually annotated by two observers. Following [8, 9], we use the annotations provided by the first observer as the ground truth. FOV masks from [8] are used as they do not come with the original dataset. As the dataset does not come with any predefined training and testing sets, following [10], the evaluation is performed using leave one-out cross validation.

4.2. Network Architecture

We use U-Net architecture as the baseline [11]. The proposed Phase U-Net architecture is illustrated in Fig. 2. The difference between the baseline U-Net architecture and the proposed Phase U-Net architecture is that the first 3×3 convolutional layer in U-Net (just after the input layer) is replaced with the ReLPU layer (with output channel of size 512) in the Phase U-Net. The rest of the network is same consisting of a contracting path and an expansive path with skip connections. It consists of repeated application of the standard 3×3 unpadded convolutions, each followed by the ReLU activation function. For downsampling, 2×2 max pooling with stride 2 is used. After each downsampling step, the number of feature channels are doubled. Every step in the expansive path consists of an upsampling of the feature map followed by a 2×2 convolution that halves the number of feature channels, a concatenation with the corresponding feature map from the contracting path, and two standard 3×3 convolutional layers, each followed by the ReLU activation function. The final layer is a 1×1 convolution followed by the softmax activation function which is used to map each 32 component feature vector to the desired number of classes which is 2 in our case. In total, the network has 9 convolutional layers with 1 ReLPU layer.

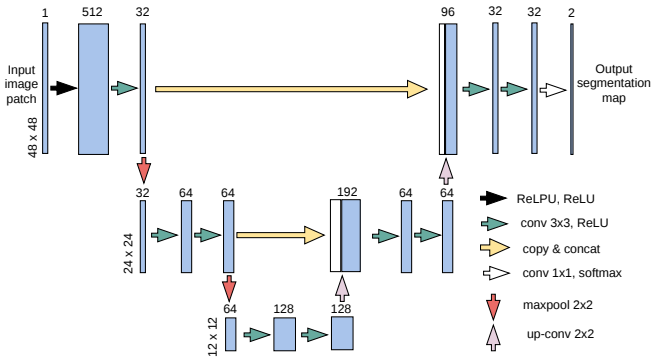


Fig. 2: The proposed phase U-Net architecture for structured prediction designed to segment retinal blood vessels from given fundus image.

	AUC ROC	
Method	**DRIVE**	**STARE**
Soares *et al.* [12]	96.14	96.71
Azzopardi *et al.* [9]	96.14	95.63
Osareh *et al.* [13]	96.50	-
Roychowdhury *et al.* [14]	96.70	96.88
Fraz *et al.* [15]	97.47	97.68
Qiaoliang *et al.* [16]	97.38	98.79
Melinscak *et al.* [17]	97.49	-
Liskowski *et al.* [10]	97.90	99.28
Dasgupta *et al.* [18]	97.44	-
Yan *et al.* [19]	97.52	98.01
Wu *et al.* [20]	98.07	-
Zhang *et al.* [21]	97.99	98.82
U-Net [11] (baseline)	97.90	98.92
Phase U-Net (ours)	**98.31**	**99.30**

Table 1: Performance results compared to other state-of-the-art methods on DRIVE and STARE datasets in terms of area under the ROC curve.

4.3. Training

We first pre-process the images by converting them to the gray-scale format and normalize them by subtracting the

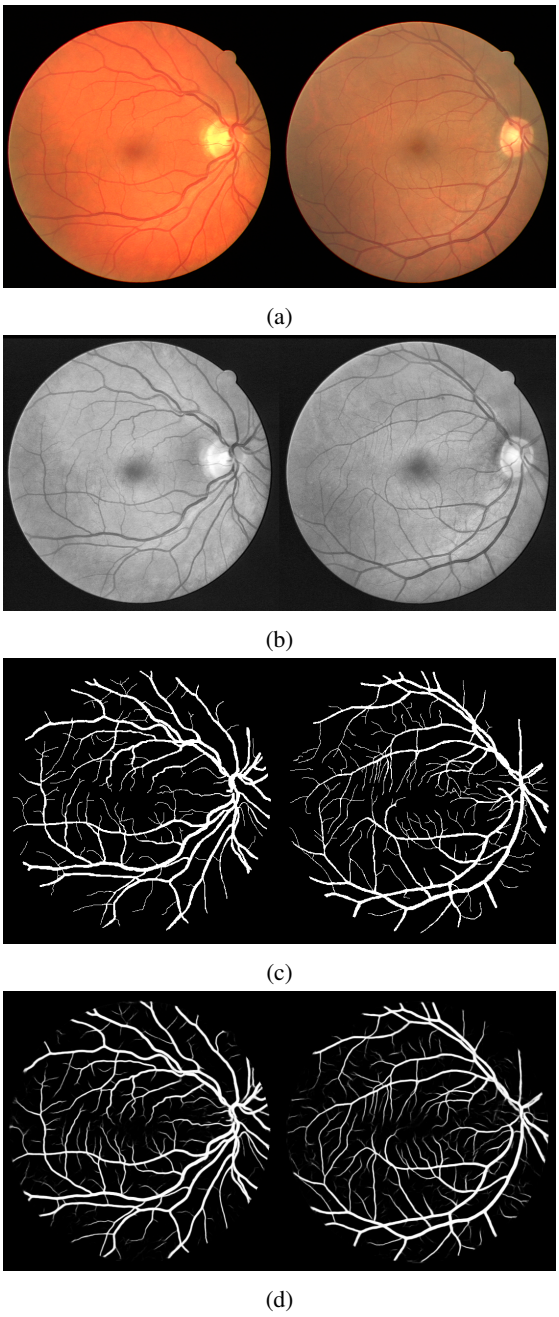
mean and dividing by the standard deviation of its elements (independently in the R, G, and B channels). Next, contrast limited adaptive histogram equalization (CLAHE) [22] and gamma adjustment are applied on the normalized images. Finally, the intensity values are scaled to have a minimum value of 0 and a maximum value of 1 to obtain the pre-processed images.

The training of the proposed model is performed on sub-images (patches) of the pre-processed full images. A total of 9500 patches, each of dimension 48×48, are obtained by randomly selecting their centers inside the full image. Further, the patches partially or completely outside the Field Of View (FOV) are selected. In this way, the model learns how to discriminate the FOV border from the blood vessels. During training, the patches were randomly flipped, rotated, shifted, and applied noise.

Our models are implemented using Keras with TensorFlow as background on a system with i7-8700 processor, 32 Gb RAM, and a single Nvidia Titan Xp 12 Gb GPU. It was optimized using Stochastic Gradient Descent (SGD) optimizer with a momentum value of 0.9, crossentropy as loss, and a batch size (of patches) of 32. All the weights are initialized with the orthogonal initializer. We start with a learning rate of 0.01 and decrease it to 0.0001 over the course of training. We train the proposed network on the DRIVE dataset from scratch and use the obtained trained model to do transfer learning on the STARE dataset.

4.4. Results

Table 1 presents the quantitative comparison of the proposed network with the previous works. Following [10], we use area under the ROC curve as the metric to evaluate our model. Note that the proposed Phase U-Net architecture significantly improves the accuracy of the baseline U-Net architecture, thus proving that the local phase information extracted from the input image has better representation capability than just the spatial information which is used by the baseline U-Net architecture. Fig. 3 shows the visualizations of the segmented map that is output by our network along with the corresponding input image and the ground truth map. We observe qualitatively that the predicted segmented map appears very adjacent to the ground truth map.



(a)

(b)

(c)

(d)

Fig. 3: Visualization of the prediction made by our proposed model on two samples randomly taken from the DRIVE dataset: (a) Original images, (b) Pre-processed images, (c) Corresponding ground truths, and (d) Segmented outputs.

5. CONCLUSION

This work explores the application of the local phase information in CNNs especially in medical image segmentation networks such as U-Net. Specifically, we propose ReLPU, an efficient and trainable local phase-base convolutional layer. The ReLPU layer when used at the top (just after the input layer) of the segmentation network U-Net such that the input to the network is the local phase information (rectified) of the input images, significantly improves the performance of the baseline U-Net model. We show this on the task of segmenting blood vessels in fundus images of two standard datasets, DRIVE and STARE, achieving state-of-the-art results. Future work could involve more analysis of the ReLPU layer in order to improve its performance and applying it to applications and scenario where it can be useful.

6. REFERENCES

[1] B Hinman, Jared Bernstein, and D Staelin, "Short-space fourier transform image processing," in *ICASSP*, 1984, vol. 9, pp. 166–169.

[2] Ville Ojansivu and Janne Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International conference on image and signal processing*, 2008, pp. 236–243.

[3] Yuan-Shan Lee, Chien-Yao Wang, Shu-Fan Wang, Jia-Ching Wang, and Chung-Hsien Wu, "Fully complex deep neural network for phase-incorporating monaural source separation," in *ICASSP*, 2017, pp. 281–285.

[4] Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra, "Reducing overfitting in deep networks by decorrelating representations," *ICLR*, 2016.

[5] Wei Xiong, Bo Du, Lefei Zhang, Ruimin Hu, and Dacheng Tao, "Regularizing deep convolutional neural networks with a structured decorrelation constraint.," in *ICDM*, 2016, pp. 519–528.

[6] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[7] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.

[8] Diego Marín, Arturo Aquino, Manuel Emilio Gegúndez-Arias, and José Manuel Bravo, "A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features," *IEEE Transactions on Medical Imaging*, vol. 30, no. 1, pp. 146, 2011.

[9] George Azzopardi, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov, "Trainable cosfire filters for vessel delineation with application to retinal images," *Medical Image Analysis*, vol. 19, no. 1, pp. 46–57, 2015.

[10] Paweł Liskowski and Krzysztof Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2369–2380, 2016.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[12] João VB Soares, Jorge JG Leandro, Roberto M Cesar, Herbert F Jelinek, and Michael J Cree, "Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification," *IEEE Transactions on Medical Imaging*, vol. 25, no. 9, pp. 1214–1222, 2006.

[13] A Osareh and B Shadgar, "Automatic blood vessel segmentation in color images of retina," *Iranian Journal of Science and Technology*, vol. 33, no. B2, pp. 191, 2009.

[14] Sohini Roychowdhury, Dara D Koozekanani, and Keshab K Parhi, "Blood vessel segmentation of fundus images by major vessel extraction and subimage classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1118–1128, 2015.

[15] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2538–2548, 2012.

[16] Qiaoliang Li, Bowei Feng, LinPei Xie, Ping Liang, Huisheng Zhang, and Tianfu Wang, "A cross-modality learning approach for vessel segmentation in retinal images.," *IEEE Transactions of Medical Imaging*, vol. 35, no. 1, pp. 109–118, 2016.

[17] Martina Melinščak, Pavle Prentašić, and Sven Lončarić, "Retinal vessel segmentation using deep neural networks," in *VISAPP*, 2015.

[18] Avijit Dasgupta and Sonam Singh, "A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation," in *ISBI*, 2017, pp. 248–251.

[19] Zengqiang Yan, Xin Yang, and Kwang-Ting Tim Cheng, "Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation," *IEEE Transactions on Biomedical Engineering*, 2018.

[20] Yicheng Wu, Yong Xia, Yang Song, Yanning Zhang, and Weidong Cai, "Multiscale network followed network model for retinal vessel segmentation," in *MICCAI*, 2018, pp. 119–126.

[21] Yishuo Zhang and Albert CS Chung, "Deep supervision with additional labels for retinal vessel segmentation task," in *MICCAI*, 2018, pp. 83–91.

[22] Stephen M Pizer, R Eugene Johnston, James P Ericksen, Bonnie C Yankaskas, and Keith E Muller, "Contrast-limited adaptive histogram equalization: speed and effectiveness," in *Visualization in Biomedical Computing*, 1990, pp. 337–345.